Discussion on EDRN Informatics Goals and Progress

JPL

August 21, 2008

Outline

- Informatics goals and their specifics
- Current Status
- Gaps
- Major milestones
- Addressing the gaps
- Progress in informatics
- Proposed items for discussion

EDRN Informatics Goals and Principles

- Develop a knowledge system that **links** together EDRN data assets into a virtual data system based on **common data elements**
- Establish an EDRN bioinformatics program that promotes the use of a **common informatics infrastructure** by EDRN sites.
- Provide an infrastructure for **capturing** EDRN biomarkers and validation study results and a mechanism for **distribution**
- Define data and software **standards** for EDRN informatics systems
- **Collaborate** with both EDRN and non-EDRN sites on informatics.
- Enable tools that support **scientific inquiry** both within and across databases and data sets.

Breaking Down the Goals…

- Develop a knowledge system that **links** together EDRN data assets into a virtual data system based on **common data elements**
  - Continue to develop and define the common data elements and the CDE Repository
  - Continue to develop systems using the common data elements to create an Integrated Knowledge Environment.
  - Ensure the common data elements form a conceptual model for biomarker research (e.g. ontology)

- Establish an EDRN Bioinformatics program that promotes the use of a **common informatics infrastructure** by EDRN sites.
  - Continue to develop and define the ontology
  - Continue to development of an online ontology that links with the CDE repository for use throughout EDRN
  - Continue to develop systems using the ontology
  - Release the ontology and data standard for wider use (e.g., working with the AACR-FDA-NCI group)
  - Deploy standard services for data sharing

Breaking Down the Goals… (cont…)

- Provide an infrastructure for **capturing** EDRN biomarkers and validation study results and a mechanism for **distribution**
  - *VSIMS* provides the validation study management support
  - *eSIS* provides annotation of studies
  - *Biomarker Database* captures the annotations and associations (biomarker/organ/study)
  - *eCAS* provides annotation, capture and distribution of scientific data sets
  - The public portal, specifically EKE, provides a means for general access and distribution

- Define data and software **standards** for EDRN informatics systems
  - The CDEs and ontology is a good start
  - RDF will help from biomarker and eSIS
  - OODT product servers help to for access to specimen information
  - But, EDRN needs a stronger presence in this area

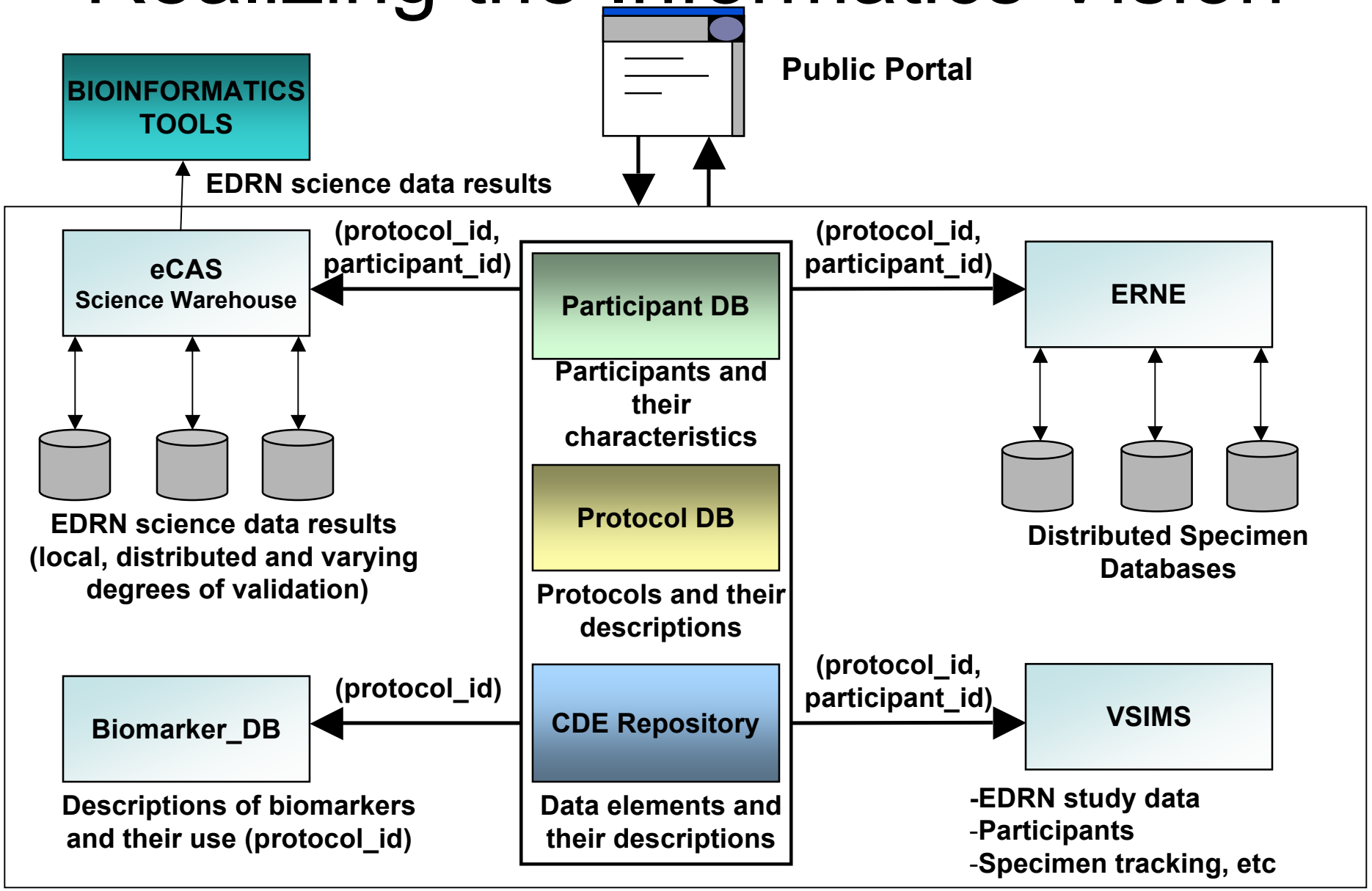
Breaking Down the Goals… (cont…)

- **Collaborate** with both EDRN and non-EDRN sites on informatics.
  - We've done some of this, but I think we can do more
  - We need another EDRN-level informatics meeting once EKE is deployed
  - We need to work with NCI to help us making connections outside EDRN

- Develop a public portal that provides **information dissemination** about EDRN programs and progress.
  - This goal has been met with the public portal. Although, content-wise, it can be improved. Some of this will be addressed through the biomarker data management activity.
  - This goal does not mention much about public access to science information and research. EKE provides that.

- Enable tools that support **scientific inquiry** both within and across databases and data sets.
  - EKE provides information integration and access, but not scientific inquiry
  - Some efforts underway with Biomarker Atlas
  - But, should other items be on our roadmap?

Realizing the Informatics Vision

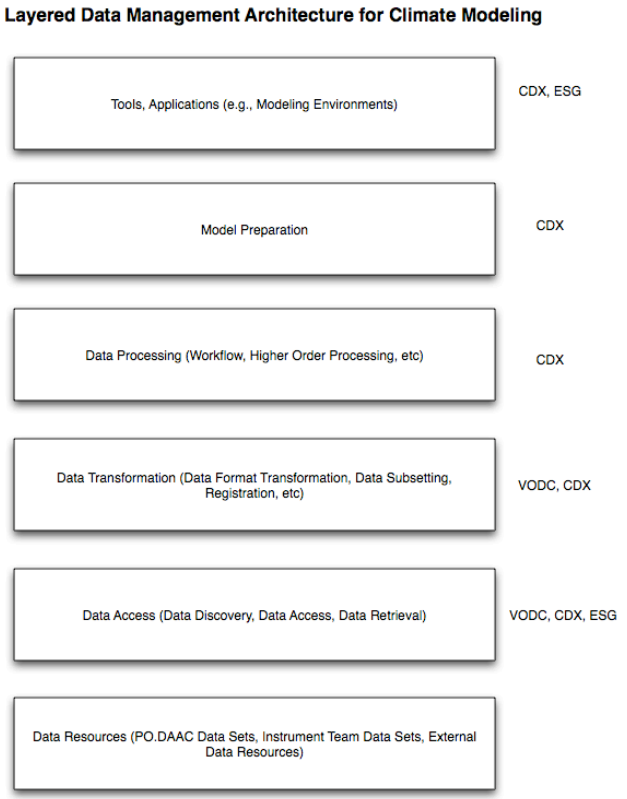


Current Status

- JPL has a new operational server called cancer.jpl.nasa.gov which is live
- Deliveries are being made to the server with a bi-weekly delivery schedule for
  - Biomarker Database
    - (http://cancer.jpl.nasa.gov/bmdb)
  - EKE
    - (http://cancer.jpl.nasa.gov/portal)
  - eCAS
    - (http://cancer.jpl.nasa.gov/ecas)
- Our operational server also uses LDAP to allow for authentication and authorization from multiple applications
  - Some challenges in getting Terpsys to install and get software ready. Long term plan is to keep LDAP, BMDB, EKE and eCAS at JPL and provide a simple shell to NCI to connect with us.
- Our development projects and pilots sit on an internal machine and include other activities such as the Biomarker Atlas for the Lung
- Dartmouth has begun working with JPL and NCI to enter biomarker information into the biomarker database

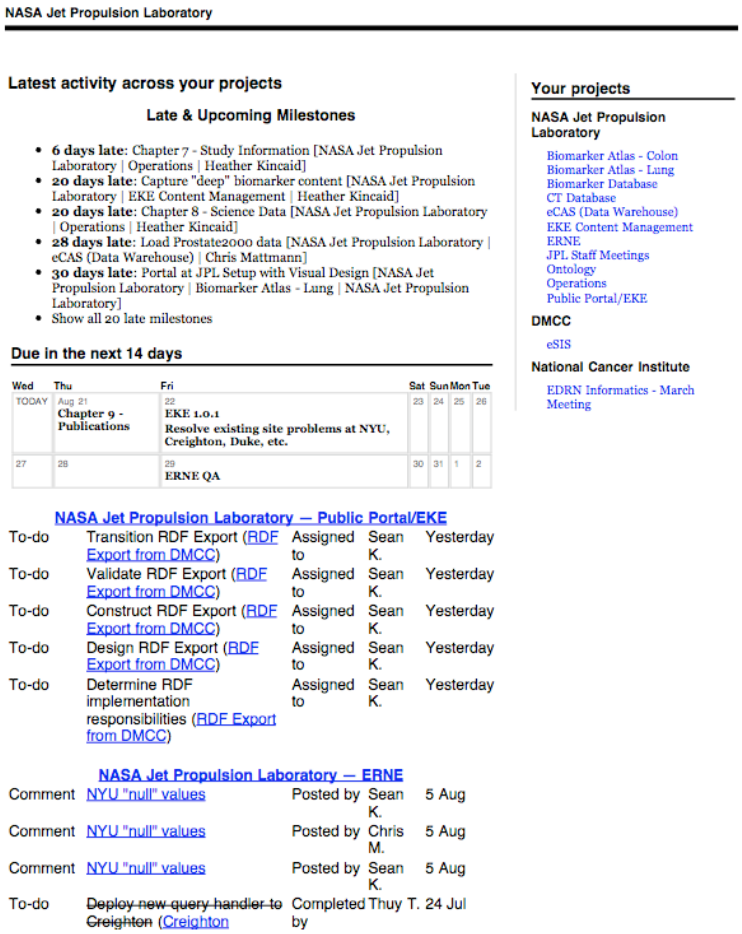
Gaps: What are we missing?

- Operational aspects of managing biomarker and eCAS databases
  - *Biomarker Data Management* needs to be coordinated from a high level EDRN perspective
- Some technical items
  - EDRN-wide IT security architecture. caBIG will be pushing this from their perspective.
  - Leveraging semantic-based technologies and search approaches
  - Tools that can plug into this environment and use the data
    - Although, in planetary and earth, a major comment that JPL often gets from the science community is not to build monolithic systems that tries to do everything. The idea is to separate the layers (see right)



Major Milestones

- Augment existing public portal so it is integrated with EKE (ASAP).
  - Deploy operationally with password and include biomarkers, science data and study information
- Finalize and load initial set of biomarkers and call for peer review (Fall 2008)
- Establish organ-specific Biomarker Data Management centers (2009)



NASA Jet Propulsion Laboratory

Latest activity across your projects

Late & Upcoming Milestones

- **6 days late**: Chapter 7 - Study Information [NASA Jet Propulsion Laboratory | Operations | Heather Kincaid]
- **20 days late**: Capture "deep" biomarker content [NASA Jet Propulsion Laboratory | EKE Content Management | Heather Kincaid]
- **20 days late**: Chapter 8 - Science Data [NASA Jet Propulsion Laboratory | Operations | Heather Kincaid]
- **28 days late**: Load Prostate2000 data [NASA Jet Propulsion Laboratory | eCAS (Data Warehouse) | Chris Mattmann]
- **30 days late**: Portal at JPL Setup with Visual Design [NASA Jet Propulsion Laboratory | Biomarker Atlas - Lung | NASA Jet Propulsion Laboratory]
- Show all 20 late milestones

Due in the next 14 days

Wed	Thu	Fri	Sat	Sun	Mon	Tue
TODAY	Aug 21 **Chapter 9 - Publications**	22 **EKE 1.0.1** **Resolve existing site problems at NYU, Creighton, Duke, etc.**	23	24	25	26
27	28	29 **ERNE QA**	30	31	1	2

Your projects

NASA Jet Propulsion Laboratory
- Biomarker Atlas - Colon
- Biomarker Atlas - Lung
- Biomarker Database
- CT Database
- eCAS (Data Warehouse)
- EKE Content Management
- ERNE
- JPL Staff Meetings
- Ontology
- Operations
- Public Portal/EKE

DMCC
- eSIS

National Cancer Institute
- EDRN Informatics - March Meeting

NASA Jet Propulsion Laboratory — Public Portal/EKE

To-do	Transition RDF Export (RDF Export from DMCC)	Assigned to	Sean K.	Yesterday
To-do	Validate RDF Export (RDF Export from DMCC)	Assigned to	Sean K.	Yesterday
To-do	Construct RDF Export (RDF Export from DMCC)	Assigned to	Sean K.	Yesterday
To-do	Design RDF Export (RDF Export from DMCC)	Assigned to	Sean K.	Yesterday
To-do	Determine RDF implementation responsibilities (RDF Export from DMCC)	Assigned to	Sean K.	Yesterday

NASA Jet Propulsion Laboratory — ERNE

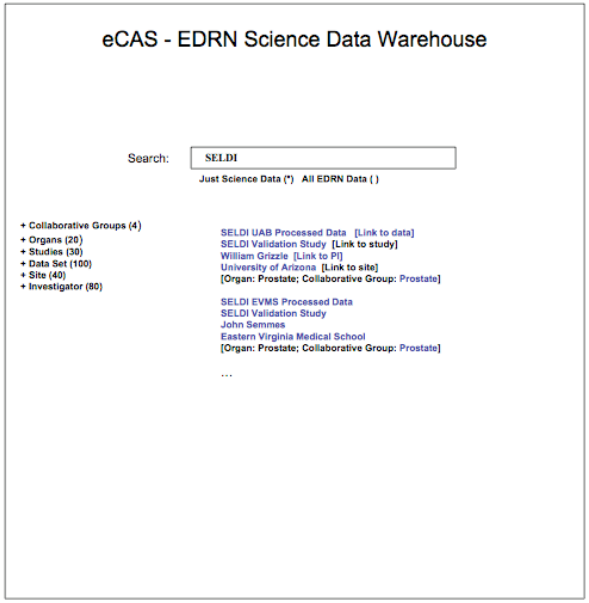
Comment	NYU "null" values	Posted by	Sean K.	5 Aug
Comment	NYU "null" values	Posted by	Chris M.	5 Aug
Comment	NYU "null" values	Posted by	Sean K.	5 Aug
To-do	~~Deploy new query handler to Creighton~~ (Creighton	Completed by	Thuy T.	24 Jul

Other Items: Security

- LDAP: Part of a common security architecture
  - Enables multiple applications to authenticate against a single, network-based service (e.g., the NCI portal, eCAS, BMDB, etc)
  - Maintains multiple roles per user
  - Applications can then implement role-based authorization
    - Collaborative groups can then see tailored views of data in EKE (e.g., biomarkers or data could be shared only with those groups)
    - NCI can have special “features” that is only available to them
  - Can be used as a shared service across EDRN

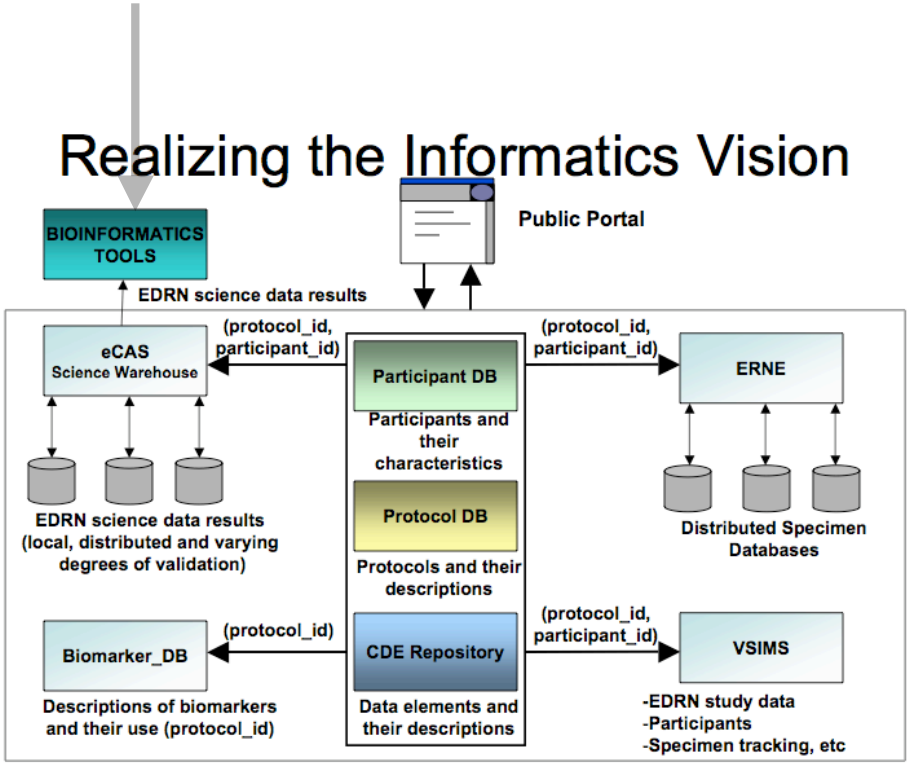
Other items: semantic based technologies

- Semantic-based technologies
  - EDRN is really building a semantic web as Mark said at the last workshop in March
  - Building a semantic based architecture allows for new types of data to be collected without requiring major changes to the infrastructure
  - Users are able to access and navigate information by following the relationships in the data
  - We see three canonical user interface approaches
    - Forms: Older approach
    - Free Text: Google approach
    - Faceted-based: Emerging approach which can really leverage relationships between data



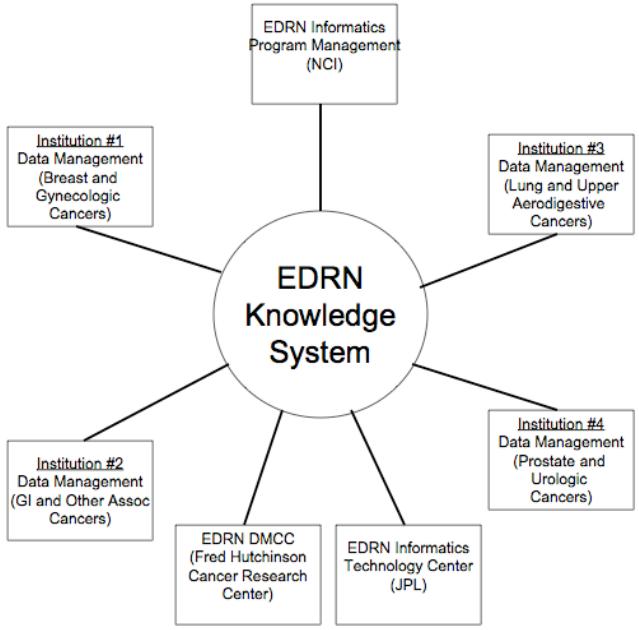
Other items: Bioinformatics tools that leverage the data

- As EDRN moves out on the infrastructure, what types of tools should or can be developed?
  - Biomarker Atlas
  - Others? Statistical tools?



Other items: Biomarker Data Management Operations

- Coordinate between NCI, JPL, DMCC and a curator
- Connect to the scientists/PIs
- Organize the curation effort
  - Recommend by organ groups
- Support relation of data across multiple systems/databases
- Support capture, review and QA of the data



Progress in cancer research towards science-driven informatics architectures

- Recognition of how to architect science-driven distributed software systems*
  - Separate the architecture into core pieces (process, data and software)
  - The “information model” is critical
    - Should provide a generalized mechanism to describe and organize data
    - Model-driven systems provide the agility to support multi-project, multi-center studies
  - Develop modular software components that can be configured based on the “information model”
    - Modularity helps to drive both longevity and agility in system designs
  - Allow for geographically distributed software components to communicate based on standards
  - Identify and implement core scientific “use cases” that help to evolve the system
  - EDRN has demonstrated this architecture can work in managing and sharing specimen information
  - JPL has done this for planetary science and is now working with international space agencies to provide access to scientific data results returned from international missions

*D. Crichton, S. Kelly, C. Mattmann, Q. Xiao, J. S. Hughes, J. Oh, M. Thornquist, D. Johnsey, S. Srivastava, L. Esserman, W. Bigbee. **A Distributed Information Services Architecture to Support Biomarker Discovery in Early Detection of Cancer.** In Proceedings of the *2nd IEEE International Conference on e-Science and Grid Computing, pp. 44*, Amsterdam, the Netherlands, December 4th- 6th, 2006.

Open Issues for Discussion

- What role should the CDEs play in biomarker data management?
- What types of bioinformatics tools are needed?
- How do we interface with the PIs?
- What types of services/support should we give to other EDRN groups that are building tools and collecting data?
- How should we engage/work with the collaborative groups?